Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction

Richard Zhang Phillip Isola Alexei A. Efros

Berkeley AI Research (BAIR) Laboratory
University of California, Berkeley
{rich.zhang,isola,efros}@eecs.berkeley.edu

Abstract

We propose split-brain autoencoders, a straightforward modification of the traditional autoencoder architecture, for unsupervised representation learning. The method adds a split to the network, resulting in two disjoint sub-networks. Each sub-network is trained to perform a difficult task – predicting one subset of the data channels from another. Together, the sub-networks extract features from the entire input signal. By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks. This method achieves state-of-the-art performance on several large-scale transfer learning benchmarks.

1. Introduction

A goal of unsupervised learning is to model raw data without the use of labels, in a manner which produces a useful representation. By "useful" we mean a representation that should be easily adaptable for other tasks, unknown during training time. Unsupervised deep methods typically induce representations by training a network to solve an auxiliary or "pretext" task, such as the image *reconstruction* objective in a traditional autoencoder model, as shown on Figure 1(top). We instead force the network to solve complementary *prediction* tasks by adding a split in the architecture, shown in Figure 1 (bottom), dramatically improving transfer performance.

Despite their popularity, autoencoders have actually not been shown to produce strong representations for transfer tasks in practice [43, 34]. Why is this? One reason might be the mechanism for forcing model abstraction. To prevent a trivial identity mapping from being learned, a bottleneck is typically built into the autoencoder representation. However, an inherent tension is at play: the smaller the bottleneck, the greater the forced abstraction, but the smaller the information content that can be expressed.

Instead of forcing abstraction through compression, via a bottleneck in the network architecture, recent work has explored withholding parts of the input during training

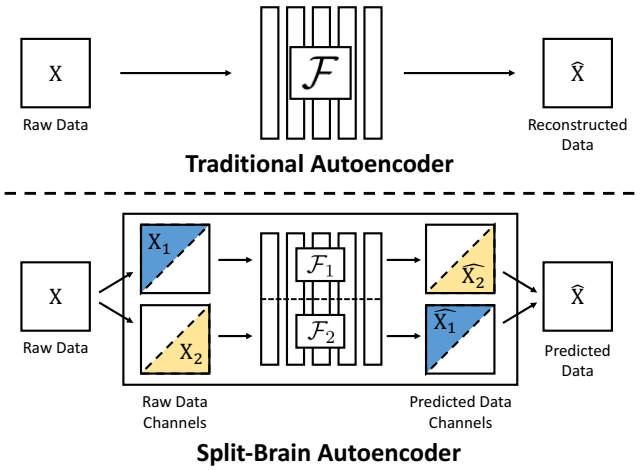


Figure 1: **Traditional vs Split-Brain Autoencoder architectures. (top)** Autoencoders learn feature representation $\mathcal{F}$ by learning to reconstruct input data $\mathbf{X}$. **(bottom)** The proposed split-brain autoencoder is composed of two disjoint sub-networks $\mathcal{F}_1, \mathcal{F}_2$, each trained to predict one data subset from another, changing the problem from reconstruction to prediction. The split-brain representation $\mathcal{F}$ is formed by concatenating the two sub-networks, and achieves strong transfer learning performance. The model is publicly available on `https://richzhang.github.io/splitbrainauto`.

[43, 34, 47]. For example, Vincent et al. [43] propose denoising autoencoders, trained to remove *iid* noise added to the input. Pathak et al. [34] propose *context encoders*, which learn features by training to inpaint large, random contiguous blocks of pixels. Rather than dropping data in the spatial direction, several works have dropped data in the channel direction, e.g. predicting color channels from grayscale (the colorization task) [26, 47].

Context encoders, while an improvement over autoencoders, demonstrate lower performance than competitors on large-scale semantic representation learning benchmarks [47]. This may be due to several reasons. First, im-

	auxiliary task type	domain gap	input handicap
Autoencoder [19]	reconstruction	no	no
Denoising autoencoder [43]	reconstruction	suffers	no
Context Encoder [34]	prediction	no	suffers
Cross-Channel Encoder [47, 27]	prediction	no	suffers
Split-Brain Autoencoder	**prediction**	**no**	**no**

Table 1: **Qualitative Comparison** We summarize various qualitative aspects inherent in several representation learning techniques. **Auxiliary task type:** pretext task predicated on reconstruction or prediction. **Domain gap:** gap between the input data during unsupervised pre-training and testing time. **Input handicap:** input data is systematically dropped out during test time.

age synthesis tasks are known to be notoriously difficult to evaluate [35] and the loss function used in [34] may not properly capture inpainting quality. Second, the model is trained on images with missing chunks, but applied, at test time, to full images. This causes a "domain gap" between training and deployment. Third, it could simply be that the inpainting task in [34] could be adequately solved without high-level reasoning, instead mostly just copying low and mid-level structure from the surround.

On the other hand, colorization turns out to be a surprisingly effective pretext task for inducing strong feature representations [47, 27]. Though colorization, like inpainting, is a synthesis task, the spatial correspondence between the input and output pairs may enable basic off-the-shelf loss functions to be effective. In addition, the *systematic*, rather than *stochastic* nature of the input corruption removes the pre-training and testing domain gap. Finally, while inpainting may admit reasoning mainly about textural structure, predicting accurate color, e.g., knowing to paint a schoolbus yellow, may more strictly require object-level reasoning and therefore induce stronger semantic representations. Colorization is an example of what we refer to as a *cross-channel encoding* objective, a task which directly predicts one subset of data channels from another.

In this work, we further explore the space of cross-channel encoders by systematically evaluating various channel translation problems and training objectives. Cross-channel encoders, however, face an inherent handicap: different channels of the input data are not treated equally, as part of the data is used for feature extraction and another as the prediction target. In the case of colorization, the network can only extract features from the grayscale image and is blind to color, leaving the color information unused. A qualitative comparison of the different methods, along with their inherent strengths and weaknesses, is summarized in Table 1.

Might there be a way to take advantage of the underlying principle of cross-channel encoders, while being able to extract features from the entire input signal? We propose an architectural modification to the autoencoder paradigm: adding a single split in the network, resulting in two disjoint, concatenated, sub-networks. Each sub-network is trained as a cross-channel encoder, predicting one subset of channels of the input from the other. A variety of auxiliary cross-channel prediction tasks may be used, such as colorization and depth prediction. For example, on RGB images, one sub-network can solve the problem of colorization (predicting a and b channels from the L channel in Lab colorspace), and the other can perform the opposite (synthesizing L from a, b channels). In the RGB-D domain, one sub-network may predict depth from images, while the other predicts images from depth. The architectural change induces the same forced abstraction as observed in cross-channel encoders, but is able to extract features from the full input tensor, leaving nothing on the table.

Our contributions are as follows:

- We propose the split-brain autoencoder, which is composed of concatenated cross-channel encoders, trained using *raw data* as its own supervisory signal.
- We demonstrate state-of-the-art performance on several semantic representation learning benchmarks in the RGB and RGB-D domains.
- To gain a better understanding, we perform extensive ablation studies by (i) investigating cross-channel prediction problems and loss functions and (ii) researching alternative aggregation methods for combining cross-channel encoders.

2. Related Work

Many unsupervised learning methods have focused on modeling raw data using a reconstruction objective. Autoencoders [19] train a network to reconstruct an input image, using a representation bottleneck to force abstraction. Denoising autoencoders [43] train a network to undo a random *iid* corruption. Techniques for modeling the probability distribution of images in deep frameworks have also been explored. For example, variational autoencoders (VAEs) [23] employ a variational Bayesian approach to modeling the data distribution. Other probabilistic models include restricted Boltzmann machines (RBMs) [40], deep Boltzmann machines (DBMs) [37], generative adversarial networks (GANs) [15], autoregressive models (Pixel-RNN [42] and Pixel-CNN [31]), bidirectional GANs (BiGANs) [8] and Adversarially Learned Inference (ALI) [9], and real NVP [6]. Many of these methods [19, 43, 8, 9, 37] have been evaluated for representation learning.

Another form of unsupervised learning, sometimes referred to as "self-supervised" learning [4], has recently grown in popularity. Rather than predicting labels annotated by humans, these methods predict pseudo-labels computed from the raw data itself. For example, image colorization [47, 26] has been shown to be an effective

pretext task. Other methods generate pseudo-labels from egomotion [1, 22], video [45, 29], inpainting [34], co-occurence [21], context [7, 30], and sound [32, 5, 4]. Concurrently, Pathak et al. [33] use motion masks extracted from video data. Also in these proceedings, Larsson et al. [27] provide an in-depth analysis of colorization for self-supervision. These methods generally focus on a single supervisory signal and involve some engineering effort. In this work, we show that simply predicting raw data channels with standard loss functions is surprisingly effective, often outperforming previously proposed methods.

The idea of learning representations from multisensory signals also shows up in structure learning [2], co-training [3], and multi-view learning [46]. Our method is especially related to [4, 5, 41], which use bidirectional data prediction to learn representations from two sensory modalities.

A large body of additional work in computer vision and graphics focuses on image channel prediction as an end in itself, such as colorization [47, 26, 20], depth prediction [10], and surface normal prediction [10, 44]. In contrast, rather than focusing on the graphics problem, we explore its utility for representation learning.

3. Methods

In Section 3.1, we define the paradigm of cross-channel encoding. In Section 3.2, we propose the split-brain autoencoder and explore alternatives methods for aggregating multiple cross-channel encoders into a single network.

3.1. Cross-Channel Encoders

We would like to learn a deep representation on input data tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, with $\mathbf{C}$ channels. We split the data into $\mathbf{X_1} \in \mathbb{R}^{H \times W \times C_1}$ and $\mathbf{X_2} \in \mathbb{R}^{H \times W \times C_2}$, where $C_1, C_2 \subseteq C$, and then train a deep representation to solve the prediction problem $\widehat{\mathbf{X_2}} = \mathcal{F}(\mathbf{X_1})$. Function $\mathcal{F}$ is learned with a CNN, which produces a layered representation of input $\mathbf{X_1}$, and we refer to each layer l as $\mathcal{F}^l$. By performing this *pretext* task of predicting $\mathbf{X_2}$ from $\mathbf{X_1}$, we hope to achieve a representation $\mathcal{F}(\mathbf{X}_1)$ which contains high-level abstractions or semantics.

This prediction task can be trained with various loss functions, and we study whether the loss function affects the quality of the learned representation. To begin, we explore the use of ℓ_2 regression, as shown in Equation 1.

$$\ell_2(\mathcal{F}(\mathbf{X_1}), \mathbf{X_2}) = \tfrac{1}{2} \sum_{h,w} \|\mathbf{X_2}_{h,w} - \mathcal{F}(\mathbf{X_1})_{h,w}\|_2^2 \quad (1)$$

We also study the use of a classification loss. Here, the target output $\mathbf{X_2} \in \mathbb{R}^{H \times W \times C_2}$ is encoded with function $\mathcal{H}$ into a *distribution* $\mathbf{Y_2} \in \mathbf{\Delta}^{H \times W \times Q}$, where Q is the number of elements in the quantized output space. Network $\mathcal{F}$ is then trained to predict a distribution, $\widehat{\mathbf{Y_2}} = \mathcal{F}(\mathbf{X_1}) \in \mathbf{\Delta}^{H \times W \times Q}$. A standard cross-entropy loss between the predicted and ground truth distributions is used, as shown Equation 2.

$$\ell_{cl}(\mathcal{F}(\mathbf{X_1}), \mathbf{X_2}) = -\sum_{h,w} \sum_{q} \mathcal{H}(\mathbf{X_2})_{h,w,q} \log(\mathcal{F}(\mathbf{X_1})_{h,w,q}) \quad (2)$$

In [47], the authors discover that classification loss is more effective for the graphics task of automatic colorization than regression. We hypothesize that for some tasks, especially those with inherent uncertainty in the prediction, the classification loss may lead to better representations as well, as the network will be incentivized to match the whole distribution, and not only predict the first moment.

Note that with input and output sets $C_1, C_2 = C$, and an ℓ_2 regression loss, the objective becomes identical to the autoencoder objective.

3.2. Split-Brain Autoencoders as Aggregated Cross-Channel Encoders

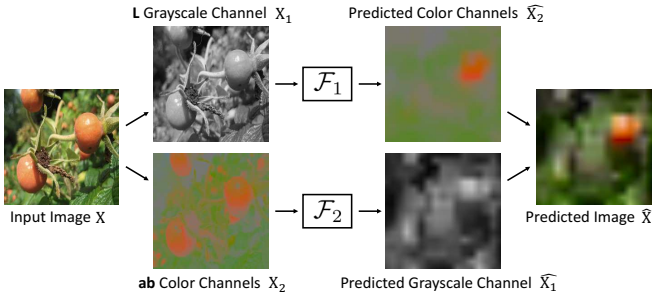
We can train multiple cross-channel encoders, $\mathcal{F}_1$, $\mathcal{F}_2$, on opposite prediction problems, with loss functions L_1, L_2, respectively, described in Equation 3.

$$\begin{aligned} \mathcal{F}_1^* &= \arg\min_{\mathcal{F}_1} L_1(\mathcal{F}_1(\mathbf{X_1}), \mathbf{X_2}) \\ \mathcal{F}_2^* &= \arg\min_{\mathcal{F}_2} L_2(\mathcal{F}_2(\mathbf{X_2}), \mathbf{X_1}) \end{aligned} \quad (3)$$

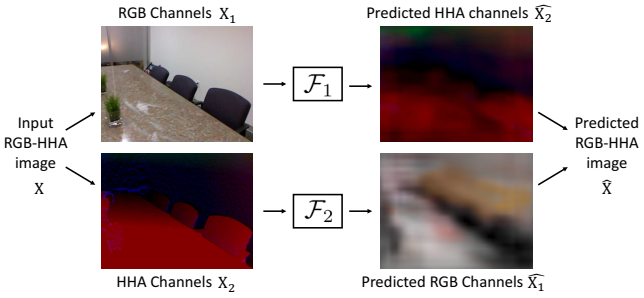
By concatenating the representations layer-wise, $\mathcal{F}^l = \{\mathcal{F}_1^l, \mathcal{F}_2^l\}$, we achieve a representation $\mathcal{F}$ which is pre-trained on full input tensor $\mathbf{X}$. Example split-brain autoencoders in the image and RGB-D domains are shown in Figures 2(a) and (b), respectively. If $\mathcal{F}$ is a CNN of a desired fixed size, e.g., AlexNet [25], we can design the sub-networks $\mathcal{F}_1, \mathcal{F}_2$ by splitting each layer of the network $\mathcal{F}$ in half, along the channel dimension. Concatenated representation $\mathcal{F}$ will then have the appropriate dimensionality, and can be simply implemented by setting the `group` parameter to 2 in most deep learning libraries. As each channel in the representation is only connected to half of the channels in the preceding layer, the number of parameters in the network is actually halved, relative to a full network.

Note that the input and the output to the network $\mathcal{F}$ is the full input $\mathbf{X}$, the same as an autoencoder. However, due to the split nature of the architecture, the network $\mathcal{F}$ is trained to *predict* $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}\}$, rather than simply *reconstruct* it from the input. In essence, an architectural change in the autoencoder framework induces the same forced abstraction achieved by cross-channel encoding.

Alternative Aggregation Technique We found the split-brain autoencoder, which aggregates cross-channel encoders through concatenation, to be more effective than several alternative strategies. As a baseline, we also explore an alternative: the same representation $\mathcal{F}$ can be trained to perform both mappings simultaneously. The loss function is described in Equation 4, with a slight abuse of notation:



(a) ***Lab* Images**



(b) **RGB-D Images**

Figure 2: **Split-Brain Autoencoders applied to various domains** (a) ***Lab* images** Input images are divided into the L channel, which contains grayscale information, and the a and b channels, which contain color information. Network $\mathcal{F}_1$ performs automatic colorization, whereas network $\mathcal{F}_2$ performs grayscale prediction. (b) **RGB-D images** Input data $\mathbf{X}$ contains registered RGB and depth images. Depth images are encoded using the HHA encoding [17]. Image representation $\mathcal{F}_1$ is trained by predicting HHA channels. Representation $\mathcal{F}_2$ on HHA images is learned by predicting images in Lab space. Note that the goal of performing these synthesis tasks is to induce representations $\mathcal{F}_1, \mathcal{F}_2$ that transfer well to other tasks.

here, we redefine $\mathbf{X_1}$ to be the same shape as original input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, with channels in set $C \backslash C_1$ zeroed out (along with the analogous modification to $\mathbf{X_2}$).

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} L_1(\mathcal{F}(\mathbf{X_1}), \mathbf{X_2}) + L_2(\mathbf{X_1}, \mathcal{F}(\mathbf{X_2})) \quad (4)$$

The network only sees data subsets but never full input $\mathbf{X}$. To alleviate this problem, we mix in the autoencoder objective, as shown in Equation 5, with $\lambda \in [0, \frac{1}{2}]$.

$$\begin{aligned}\mathcal{F}^* = \arg\min_{\mathcal{F}} \lambda L_1(\mathcal{F}(\mathbf{X_1}), \mathbf{X_2}) + \lambda L_2(\mathcal{F}(\mathbf{X_2}), \mathbf{X_1}) \\ + (1 - 2\lambda) L_3(\mathbf{X}, \mathcal{F}(\mathbf{X}))\end{aligned} \quad (5)$$

Note that unlike the split-brain architecture, in these objectives, there is a domain gap between the distribution of pretraining data and the full input tensor $\mathbf{X}$.

4. Experiments

In Section 4.1, we apply our proposed split-brain autoencoder architecture to learn unsupervised representations on large-scale image data from ImageNet [36]. We evaluate on established representation learning benchmarks and demonstrate state-of-the-art performance relative to previous unsupervised methods [24, 7, 45, 34, 32, 8, 29]. In Section 4.2, we apply the proposed method on the NYU-D dataset [38], and show performance above baseline methods.

4.1. Split-Brain Autoencoders on Images

We work with image data $\mathbf{X}$ in the *Lab* color space, and learn cross-channel encoders with $\mathbf{X}_1$ representing the L, or lightness channel, and $\mathbf{X}_2$ containing the ab channels, or color information. This is a natural choice as (i) networks such as Alexnet, trained with grouping in their architecture, naturally separate into grayscale and color [25] even in a fully-supervised setting, and (ii) the individual cross-channel prediction problem of colorization, L to ab, has produced strong representations [47, 26]. In preliminary experiments, we have also explored different cross-channel prediction problems in other color spaces, such as *RGB* and *YUV*. We found the *L* and *ab* to be most effective data split.

To enable comparisons to previous unsupervised techniques, all of our trained networks use AlexNet architectures [25]. Concurrent work from Larsson et al. [27] shows large performance improvements for the colorization task when using deeper networks, such as VGG-16 [39] and ResNet [18]. Because we are training for a pixel-prediction task, we run the network fully convolutionally [28]. Using the 1.3M ImageNet dataset [36] (without labels), we train the following aggregated cross-channel encoders:

- **Split-Brain Autoencoder (cl,cl) (Our full method)**: A split-brain autoencoder, with one half performing colorization, and the other half performing grayscale prediction. The top-level architecture is shown in Figure 2(a). Both sub-networks are trained for classification (cl), with a cross-entropy objective. (In Figure 2(a), the predicted output is a per-pixel probability distribution, but is visualized with a point estimate using the annealed-mean [47].)
- **Split-Brain Autoencoder (reg,reg)**: Same as above, with both sub-networks trained with an ℓ_2 loss (reg).
- **Ensembled L→ab**: Two concatenated disjoint subnetworks, both performing colorization (predicting ab from L). One subnetwork is trained with a classification objective, and the other with regression.
- **(L,ab)→(ab,L)**: A single network for both colorization and grayscale prediction, with regression loss, as described in Equation 4. This explores an alternative method for combining cross-channel encoders.
- **(L,ab,Lab)→(ab,L,Lab)**: $\lambda = \frac{1}{3}$ using Equation 5.

Task Generalization on ImageNet Classification [36]					
Method	**conv1**	**conv2**	**conv3**	**conv4**	**conv5**
ImageNet-labels [25]	19.3	36.3	44.2	48.3	50.5
Gaussian	11.6	17.1	16.9	16.3	14.1
Krähenbühl et al. [24]	17.5	23.0	24.5	23.2	20.6
[1]Noroozi & Favaro [30]	19.2	30.1	34.7	33.9	28.3
Doersch et al. [7]	16.2	23.3	30.2	31.7	29.6
Donahue et al. [8]	**17.7**	24.5	31.0	29.9	28.0
Pathak et al. [34]	14.1	20.7	21.0	19.8	15.5
Zhang et al. [47]	13.1	24.8	31.0	32.6	31.8
Lab→Lab	12.9	20.1	18.5	15.1	11.5
Lab(drop50)→Lab	12.1	20.4	19.7	16.1	12.3
L→ab(cl)	12.5	25.4	32.4	33.1	32.0
L→ab(reg)	12.3	23.5	29.6	31.1	30.1
ab→L(cl)	11.6	19.2	22.6	21.7	19.2
ab→L(reg)	11.5	19.4	23.5	23.9	21.7
(L,ab)→(ab,L)	15.1	22.6	24.4	23.2	21.1
(L,ab,Lab)→(ab,L,Lab)	15.4	22.9	24.0	22.0	18.9
Ensembled L→ab	11.7	23.7	30.9	32.2	31.3
Split-Brain Auto (reg,reg)	17.4	27.9	33.6	34.2	32.3
Split-Brain Auto (cl,cl)	**17.7**	**29.3**	**35.4**	**35.2**	**32.8**

Table 2: **Task Generalization on ImageNet Classification** To test unsupervised feature representations, we train linear logistic regression classifiers on top of each layer to perform 1000-way ImageNet classification, as proposed in [47]. All weights are frozen and feature maps spatially resized to be $\sim$9000 dimensions. All methods use AlexNet variants [25], and were pre-trained on ImageNet without labels, except for **ImageNet-labels**. Note that the proposed split-brain autoencoder achieves the best performance on all layers across unsupervised methods.

Single cross-channel encoders are ablations of our main method. We systematically study combinations of loss functions and cross-channel prediction problems.

- **L→ab(reg)**: Automatic colorization using an ℓ_2 loss.
- **L→ab(cl)**: Automatic colorization using a classification loss. We follow the quantization procedure proposed in [47]: the output ab space is binned into grid size 10×10, with a classification loss over the 313 bins that are within the ab gamut.
- **ab→L(reg)**: Grayscale prediction using an ℓ_2 loss.
- **ab→L(cl)**: Grayscale prediction using a classification loss. The L channel, which has values between 0 and 100, is quantized into 50 bins of size 2 and encoded.
- **Lab→Lab**: Autoencoder objective, reconstructing Lab from itself using an ℓ_2 regression loss, with the same architecture as the cross-channel encoders.
- **Lab(drop50)→Lab**: Same as above, with 50% of the input randomly dropped out during pre-training. This is similar to denoising autoencoders [43].

We compare to the following methods, which all use variants of Alexnet [25]. For additional details, refer to Table 3 in [47]. Note that one of these modifications resulted in a large deviation in feature map size[1].

[1]The method from [30] uses stride 2 instead of 4 in the `conv1` layer, resulting in 4× denser feature maps throughout all convolutional layers. While it is unclear how this change affects representational quality, experiments from Larsson et al. [27] indicate that changes in architecture can result in large changes in transfer performance, even given the same training task. The network uses the same number of parameters, but 5.6× the memory and 7.4× the run-time.

Dataset & Task Generalization on Places Classification [48]					
Method	**conv1**	**conv2**	**conv3**	**conv4**	**conv5**
Places-labels [48]	22.1	35.1	40.2	43.3	44.6
ImageNet-labels [25]	22.7	34.8	38.4	39.4	38.7
Gaussian	15.7	20.3	19.8	19.1	17.5
Krähenbühl et al. [24]	21.4	26.2	27.1	26.1	24.0
[1]Noroozi & Favaro [30]	23.0	32.1	35.5	34.8	31.3
Doersch et al. [7]	19.7	26.7	31.9	32.7	30.9
Wang & Gupta [45]	20.1	28.5	29.9	29.7	27.9
Owens et al. [32]	19.9	29.3	32.1	28.8	29.8
Donahue et al. [8]	**22.0**	28.7	31.8	31.3	29.7
Pathak et al. [34]	18.2	23.2	23.4	21.9	18.4
Zhang et al. [47]	16.0	25.7	29.6	30.3	29.7
L→ab(cl)	16.4	27.5	31.4	32.1	30.2
L→ab(reg)	16.2	26.5	30.0	30.5	29.4
ab→L(cl)	15.6	22.5	24.8	25.1	23.0
ab→L(reg)	15.9	22.8	25.6	26.2	24.9
Split-Brain Auto (cl,cl)	21.3	**30.7**	**34.0**	**34.1**	**32.5**

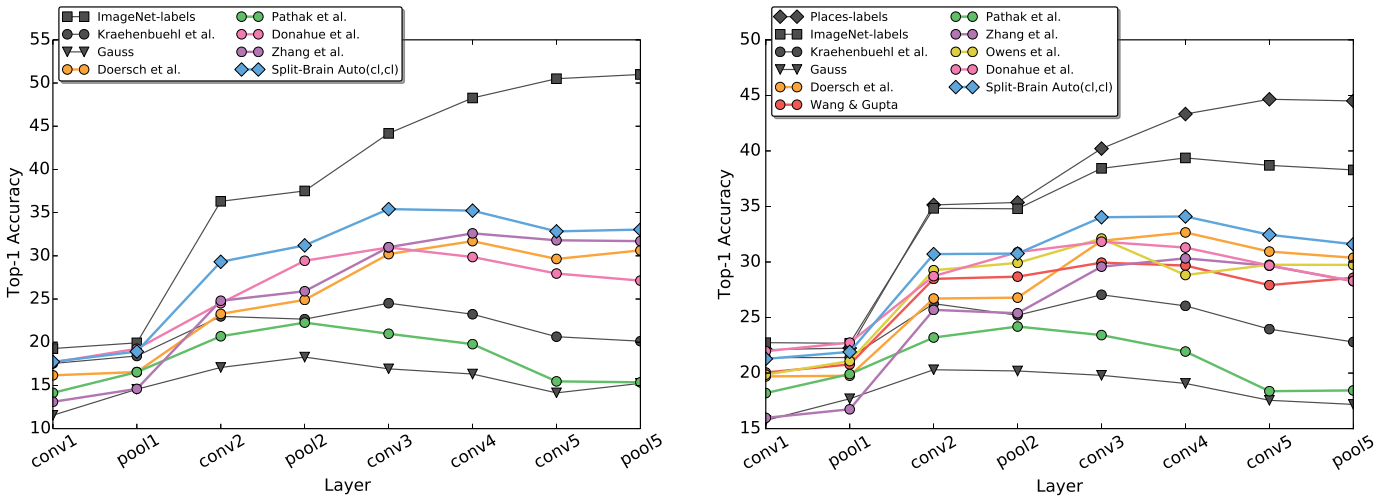
Table 3: **Dataset & Task Generalization on Places Classification** We train logistic regression classifiers on top of frozen pre-trained representations for 205-way Places classification. Note that our split-brain autoencoder achieves the best performance among unsupervised learning methods from `conv2-5` layers.

- **ImageNet-labels** [25]: Trained on ImageNet labels for the classification task in a fully supervised fashion.
- **Gaussian**: Random Gaussian initialization of weights.
- **Krähenbühl et al.** [24]: A stacked k-means initialization method.
- **Doersch et al.** [7], **Noroozi & Favaro** [30], **Pathak et al.** [34], **Donahue et al.** [8], and **Zhang et al.** [47] all pre-train on the 1.3M ImageNet dataset [36].
- **Wang & Gupta** [45] and **Owens et al.** [32] pre-train on other large-scale data.

4.1.1 Transfer Learning Tests

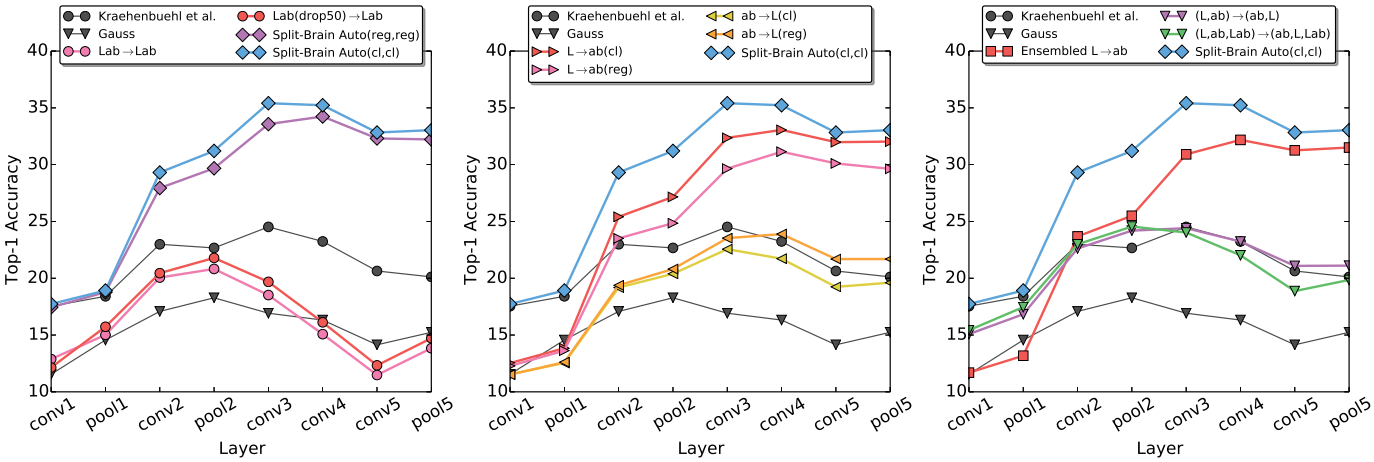
How well does the pre-text task of cross-channel prediction generalize to unseen tasks and data? We run various established large-scale representation learning benchmarks.

ImageNet [25] As proposed in [47], we test the *task* generalization of the representation by freezing the weights and training multinomial logistic regression classifiers on top of each layer to perform 1000-way ImageNet classification. Note that each classifier is a single learned linear layer, followed by a softmax. To reduce the effect of differences in feature map sizes, we spatially resize feature maps through bilinear interpolation, so that the flattened feature maps have approximately equal dimensionality (9600 for



(a) **ImageNet Classification** (b) **Places Classification**

Figure 3: **Comparison to Previous Unsupervised Methods** We compare our proposed Split-Brain Autoencoder on the tasks of (a) ImageNet classification and (b) Places Classification. Note that our method outperforms other large-scale unsupervised methods [7, 45, 34, 47, 32, 8] on all layers in ImageNet and from `conv2-5` on Places.



(a) **Autoencoder Objective** (b) **Cross-Channel Encoders** (c) **Aggregation Methods**

Figure 4: **Ablation Studies** We conduct various ablation studies on our proposed method, using the ImageNet classification benchmark proposed in [47]. Specifically, we compare (a) variations using an autoencoder objective (b) different cross-channel problems and loss functions (c) different methods for aggregating multiple cross-channel encoders.

`conv1,3,4` and 9216 for `conv2,5`). The results are shown in Table 2 and Figures 3(a) and 4.

Places [48] In the previous test, we evaluated the representation on the same input training data, the ImageNet dataset, with a different task than the pretraining tasks. To see how well the network generalizes to new input *data* as well, we run the same linear classification task on the large-scale Places dataset [48]. The dataset contains 2.4M images for training and 20.5k for validation from 205 scene categories. The results are shown in Table 3 and Figure 3(b).

PASCAL [11] To further test generalization, we fine-tune the learned representation on standard representation learning benchmarks on the PASCAL dataset, as shown in Table 4, using established testing frameworks in classification [24], detection [14], and segmentation [28]. Classification involves 20 binary classification decisions, regarding the presence or absence of 20 object classes. Detection involves drawing an accurately localized bounding box around any objects in the image, and is performed using the Fast R-CNN [14] framework. Segmentation is pixel-

Task and Data Generalization on PASCAL VOC [11]							
	Classification [24] (%mAP)			Detection [14] (%mAP)		Seg. [28] (%mIU)	
frozen layers		**conv5**	**none**		**none**		**none**
fine-tuned layers	**Ref**	**fc6-8**	**all**	**Ref**	**all**	**Ref**	**all**
ImageNet labels [25]	[47]	78.9	79.9	[24]	56.8	[28]	48.0
Gaussian	[34]	–	53.3	[34]	43.4	[34]	19.8
Autoencoder	[8]	16.0	53.8	[34]	41.9	[34]	25.2
Krähenbühl et al. [24]	[8]	39.2	56.6	[24]	45.6	[8]	32.6
Jayaraman & Grauman [22]	–	–	–	[22]	41.7	–	–
Agrawal et al. [1]	[24]	–	52.9	[24]	41.8	–	–
Agrawal et al. [1]†	[8]	31.0	54.2	[24]	43.9	–	–
Wang & Gupta [45]	[24]	–	62.8	[24]	47.4	–	–
Wang & Gupta [45]†	[24]	–	63.1	[24]	47.2	–	–
Doersch et al. [7]	[24]	–	55.3	[24]	46.6	–	–
Doersch et al. [7]†	[8]	55.1	65.3	[24]	51.1	–	–
Pathak et al. [34]	[34]	–	56.5	[34]	44.5	[34]	29.7
Donahue et al. [8]†	[8]	52.3	60.1	[8]	46.9	[8]	35.2
Misra et al. [29]	–	–	–	[29]	42.4	–	–
Owens et al. [32]	▷	54.6	54.4	[32]	44.0	–	–
Owens et al. [32]†	▷	52.3	61.3	–	–	–	–
Zhang et al. [47]†	[47]	61.5	65.9	[47]	46.9	[47]	35.6
Larsson et al. [27]◇	[27]	–	65.9	–	–	[27]	**38.4**
Pathak et al. [33]◇	[33]	–	61.0	[33]	**52.2**	–	–
Split-Brain Auto (cl,cl)†	▷	**63.0**	**67.1**	▷	46.7	▷	36.0

Table 4: **Task and Dataset Generalization on PASCAL VOC** Classification and detection on PASCAL VOC 2007 [12] and segmentation on PASCAL VOC 2012 [13], using mean average precision (mAP) and mean intersection over union (mIU) metrics for each task, with publicly available testing frameworks from [24], [14], [28]. Column **Ref** documents the source for a value obtained from a previous paper. Character ▷ indicates that value originates from this paper. †indicates that network weights have been rescaled with [24] before fine-tuning, as is common practice. Character ◇ indicates concurrent work in these proceedings.

wise labeling of the object class, either one of the 20 objects of interest or background. Here, the representation is *fine-tuned* through multiple layers of the network, rather than frozen. Prior to fine-tuning, we follow common practice and use the rescaling method from [24], which rescales the weights so that the layers learn at the same "rate", using the ratio of expected gradient magnitude over feature activation magnitude as a heuristic.

4.1.2 Split-Brain Autoencoder Performance

Our primary result is that the proposed method, **Split-Brain Auto (cl,cl)**, achieves state-of-the-art performance on almost all established self-supervision benchmarks, as seen in the last row on Tables 2, 3, 4, over previously proposed self-supervision methods, as well as our ablation baselines. Figures 3(a) and (b) shows our split brain autoencoder method compared to previous self-supervised methods [7, 45, 34, 47, 8, 32] on the ImageNet and Places classification tests, respectively. We especially note the straightforward nature of our proposed method: the network simply predicts raw data channels from other raw data channels, using a classification loss with a basic 1-hot encoding scheme.

As seen in Figure 4(a) and Table 2, the autoencoder objective by itself, **Lab→Lab**, does not lead to a strong representation. Performance is near Gaussian initialization through the initial layers, and actually falls below in the `conv5` layer. Dropping 50% of the data from the input randomly during training, **Lab(drop50)→Lab**, in the style of denoising autoencoders, adds a small performance boost of approximately 1%. A large performance boost is observed by adding a split in the architecture, **Split-Brain Auto (reg,reg)**, even with the same regression objective. This achieves 5% to 20% higher performance throughout the network, state-of-the-art compared to previous unsupervised methods. A further boost of approximately 1-2% throughout the network observed using a classification loss, **Split-Brain Auto (cl,cl)**, instead of regression.

4.1.3 Cross-Channel Encoding Objectives

Figure 4(b) compares the performance of the different cross-channel objectives we tested on the ImageNet classification benchmark. As shown in [47] and further confirmed here, colorization, **L→ab(cl)**, leads to a strong representation on classification transfer tasks, with higher performance than other unsupervised representations pre-trained on ImageNet, using inpainting [34], relative context [7], and adversarial feature networks [8] from layers from `conv2` to `pool5`. We found that the classification loss produced stronger representations than regression for colorization, consistent with the findings from concurrent work from Larsson et al. [27].

Interestingly, the task of predicting grayscale from color can also learn representations. Though colorization lends itself closely to a graphics problem, the application of grayscale prediction from color channels is less obvious. As seen in Tables 2 and 3 and Figure 4(b), grayscale prediction objectives **ab→L(cl)** and **ab→L(reg)** can learn representations above the **Gaussian** baseline. Though the learned representation by itself is weaker than other self-supervised methods, the representation is learned on a and b channels, which makes it complementary to the colorization network. For grayscale prediction, regression results in higher performance than classification. Choosing the appropriate loss function for a given channel prediction problem is an open problem. However, note that the performance difference is typically small, indicating that the cross-channel prediction problem is often times an effective method, even without careful engineering of the objective.

4.2. Split-Brain Autoencoders on RGB-D

We also test the split-brain autoencoder method on registered images and depth scans from NYU-D [38]. Because RGB and depth images are registered spatially, RGB-D data can be readily applied in our proposed framework. We split the data by modality, predicting RGB from D and vice-versa. Previous work in the video and audio domain [5]

Method	Data	Label	RGB	D	RGB-D
Gupta et al. [17]	1M ImNet [36]	✓	27.8	41.7	47.1
Gupta et al. [16]	1M ImNet [36]	✓	27.8	34.2	44.4
Gaussian	None		–	28.1	–
Krähenbühl et al. [24]	20 NYU-D [38]		12.5	32.2	34.5
Split-Brain Autoencoder	10k NYU-D [38]		**18.9**	**33.2**	**38.1**

Table 5: **Split-Brain Autoencoder Results on RGB-D images** We perform unsupervised training on 10k RGB-D keyframes from the NYU-D [38] dataset, extracted by [17]. We pre-train representations on RGB images using ℓ_2 loss on depth images in *HHA* space. We pre-train *HHA* representations on L and ab channels using ℓ_2 and classification loss, respectively. We show performance gains above Gaussian and Krähenbühl et al. [24] initialization baselines. The methods proposed by Gupta et al. [16, 17] use 1.3M labeled images for supervised pre-training. We use the test procedure from [17]: Fast R-CNN [14] networks are first trained individually in the RGB and D domains separately, and then ensembled together by averaging (RGB-D).

suggest that separating modalities, rather than mixing them, provides more effective splits. This choice also provides easy comparison to the test procedure introduced by [16].

Dataset & Detection Testbed The NYUD dataset contains 1449 RGB-D labeled images and over 400k unlabeled RGB-D video frames. We use 10k of these unlabeled frames to perform unsupervised pre-training, as extracted from [17]. We evaluate the representation on the 1449 labeled images for the detection task, using the framework proposed in [17]. The method first trains individual detectors on the RGB and D domains, using the Fast R-CNN framework [14] on an AlexNet architecture, and then late-fuses them together through ensembling.

Unsupervised Pre-training We represent depth images using the HHA encoding, introduced in [16]. To learn image representation $\mathcal{F}_{HHA}$, we train an Alexnet architecture to regress from RGB channels to HHA channels, using an ℓ_2 regression loss.

To learn depth representations, we train an Alexnet on HHA encodings, using ℓ_2 loss on *L* and classification loss on *ab* color channels. We chose this combination, as these objectives performed best for training individual cross-channel encoders in the image domain. The network extracts features up to the `conv5` layer, using an Alexnet architecture, and then splits off into specific branches for the L and ab channels. Each branch contains AlexNet-type `fc6-7` layers, but with 512 channels each, evaluated fully convolutionally for pixel prediction. The loss on the ab term was weighted 200$\times$ with respect to the L term, so the gradient magnitude on the `pool5` representation from channel-specific branches were approximately equal throughout training.

Across all methods, weights up to the `conv5` layer are copied over during fine-tuning time, and `fc6-7` layers are randomly initialized, following [16].

Results The results are shown in Table 5 for detectors learned in RGB and D domains separately, as well as the ensembled result. For a Gaussian initialization, the RGB detector did not train using default settings, while the depth detector achieved performance of 28.1%. Using the stacked k-means initialization scheme from Krähenbühl et al. [24], individual detectors on RGB and D perform at 12.5% and 32.2%, while achieving 34.5% after ensembling. Pre-training with our method reaches 18.9% and 33.2% on the individual domains, above the baselines. Our RGB-D ensembled performance was 38.1%, well above the Gaussian and Krähenbühl et al. [24] baselines. These results suggest that split-brain autoencoding is effective not just on *Lab* images, but also on RGB-D data.

5. Discussion

We present split-brain autoencoders, a method for unsupervised pre-training on large-scale data. The split-brain autoencoder contains two disjoint sub-networks, which are trained as cross-channel encoders. Each sub-network is trained to predict one subset of raw data from another. We test the proposed method on *Lab* images, and achieve state-of-the-art performance relative to previous self-supervised methods. We also demonstrate promising performance on RGB-D images. The proposed method solves some of the weaknesses of previous self-supervised methods. Specifically, the method (i) does not require a representational bottleneck for training, (ii) uses input dropout to help force abstraction in the representation, and (iii) is pre-trained on the full input data.

An interesting future direction of is exploring the concatenation of more than 2 cross-channel sub-networks. Given a fixed architecture size, e.g. AlexNet, dividing the network into N disjoint sub-networks results in each sub-network becoming smaller, less expressive, and worse at its original task. To enable fair comparisons to previous large-scale representation learning methods, we focused on learning weights for a fixed AlexNet architecture. It would also be interesting to explore the regime of fixing the sub-network size and allowing the full network to grow with additional cross-channel encoders.

Acknowledgements

We thank members of the Berkeley Artificial Intelligence Research Lab (BAIR), in particular Andrew Owens, for helpful discussions, as well as Saurabh Gupta for help with RGB-D experiments. This research was supported, in part, by Berkeley Deep Drive (BDD) sponsors, hardware donations by NVIDIA Corp and Algorithmia, an Intel research grant, NGA NURI, and NSF SMA-1514512. Thanks Obama.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

[2] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[4] V. R. de Sa. Learning classification with unlabeled data. *Advances in neural information processing systems*, pages 112–112, 1994.

[5] V. R. De Sa. Sensory modality segregation. *NIPS*, 2003.

[6] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[7] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[8] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *ICLR*, 2017.

[9] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *ICLR*, 2017.

[10] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[14] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[16] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[17] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[19] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[20] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2016)*, 35(4), 2016.

[21] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson. Learning visual groups from co-occurrences in space and time. *International Conference on Learning Representations, Workshop*, 2016.

[22] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1413–1421, 2015.

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[24] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell. Data-dependent initializations of convolutional neural networks. *International Conference on Learning Representations*, 2016.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. *European Conference on Computer Vision*, 2016.

[27] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. *CVPR*, 2017.

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[29] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[30] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *European Conference on Computer Vision*, 2016.

[31] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. *NIPS*, 2016.

[32] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016.

[33] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. *CVPR*, 2017.

[34] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[35] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala. Visual equivalence: towards a new standard for image fidelity. *ACM Transactions on Graphics (TOG)*, 26(3):76, 2007.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[37] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *AISTATS*, volume 1, page 3, 2009.

[38] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

[41] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[42] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016.

[43] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[44] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.

[45] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[46] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[47] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *European Conference on Computer Vision*, 2016.

[48] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.